Annif Theseuksessa

AMKIT-metatietoryhmän Annif-webinaari
12.11.2020
Samu Viita - Kansalliskirjasto

Annif Kansalliskirjaston ylläpitämissä julkaisuarkistoissa

- Theseus ja muut Kansalliskirjaston ylläpitämät julkaisuarkistot pohjimmiltaan samaa tekniikkaa
  - perustuvat avoimen lähdekoodin Dspace-ohjelmistoon
- Tällä hetkellä pohjana Dspace 5 koodi, jota on muokattu paljon
  - Annif-kytkentä myös osaltaan vaatinut muokkauksia koodin
    - Java- XSLT- ja JavaScript-koodia
- Kansalliskirjaston ylläpitämissä julkaisuarkistoissa Annif tällä hetkellä käytössä **Theseuksessa**, **Osuvassa** ja **Trepossa**
  - Lisää asennuksia tehdään asiakkaiden toiveiden mukaan
  - Nopea ottaa käyttöön, vaikuttaa koko arkiston tasolla syöttöjärjestykseen

Annif Theseuksessa YSO-asiasanojen syötön tukena

- Annifin käyttöönoton yhteydessä syöttöjärjestystä on muutettu, tiedoston latausvaihe on nyt ennen kuvailuvaihetta
  - Theseuksen syöttölomakkeella kuvailutiedot on aiemmin syötetty ennen tiedoston lataamista
  - Myös Dspacessa oletuksena syöttö tässä järjestyksessä
- Järjestystä vaihtamalla tiedoston teksti saadaan analysoitua ja Annifin asiasanaehdotukset mukaan kuvailuvaiheeseen
- Annif-ehdotukset koskevat kenttää **dc.subject.yso**
- Annifin ehdottamien asiasanojen valinnan rinnalla käyttäjällä on mahdollisuus syöttää asiasanoja kontrolloidusta YSO-sanastosta samaan kenttään

Theseuksen ja Annifin välinen toiminta

- Kun kokoteksti on syötetty Theseukseen, se puretaan raakatekstiksi ja ajetaan kielentunnistuksen läpi
  - Teksti ja tunnistettu kieli lähetetään Annifille analysoitavaksi Annifin REST rajapinnan kautta
  - Kielentunnistus on toteutettu Open-Source pohjaisella language-detector Java-kirjastolla
    - Tuettuna suomi, ruotsi ja englanti
  - Annif antaa vastauksena JSON-muodossa asiasanaehdotukset
- Dspace-syöttölomakkeelle generoidaan YSO-asiasanojen syöttökohtaan valintalaatikko, jossa Annifin ehdottamat asiasanat

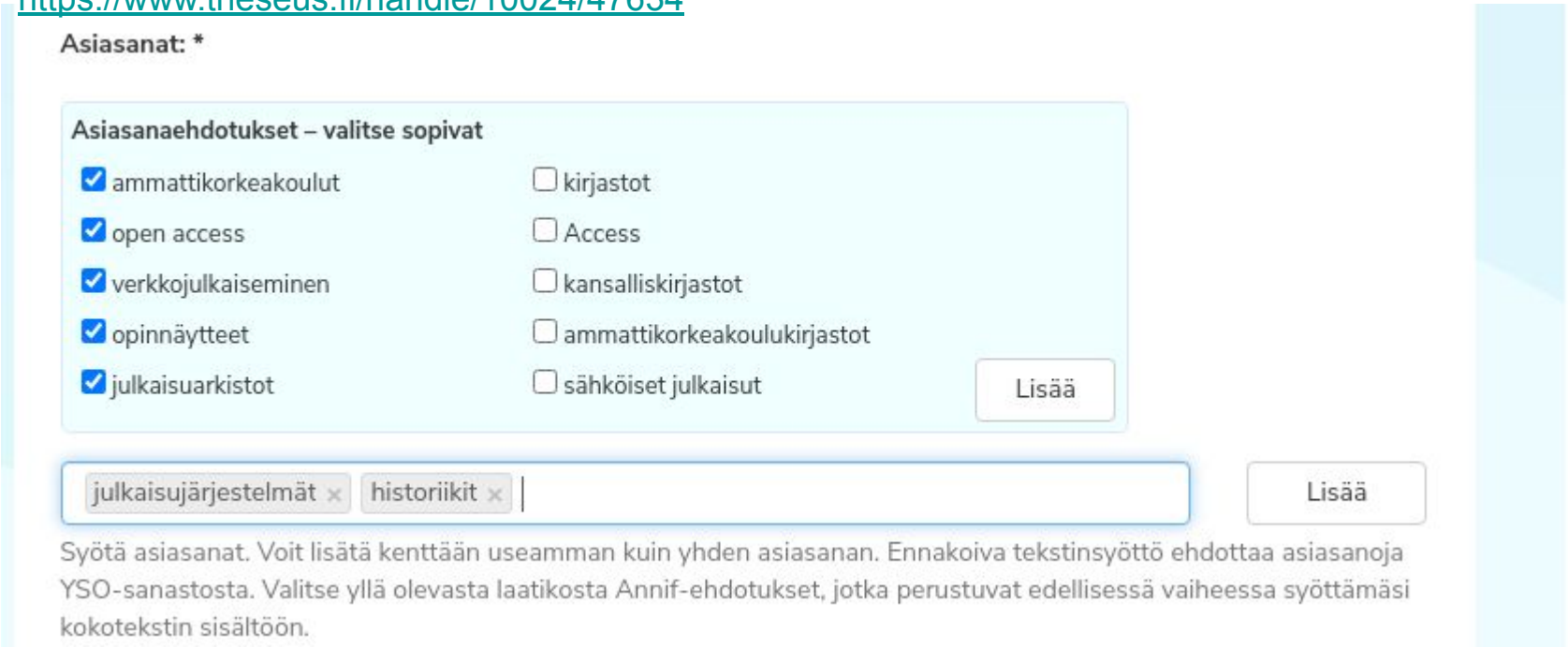
Theseuksen ja Annifin välinen toiminta

- Annif-ehdotuksia ei ole oletuksena valittuna, vaan käyttäjän pitää aktiivisesti poimia niitä YSO-asiasanojen syöttökohdassa
  - Tällä pyritään välttämään sitä, että syöttäjä hyväksyisi epähuomiossa tai kiireessä kaikki ehdotukset
- Ehdotuslaatikon alapuolella on YSO-asiasanojen syöttökenttä, johon voi kuvailla muita YSO-asiasanoja
  - Ennakoiva tekstinsyöttökenttä ehdottaa YSO-sanaston sanoja
  - Tähän hyödynnetään Finton REST-rajapintaa
- Sekä Annifin ehdotuslaatikosta poimitut, että alempaan kenttään syötetyt YSO-sanat siirtyvät valituksi kenttään dc.subject.yso

Demo Annifista Osuvan syöttölomakkeella

http://urn.fi/URN:NBN:fi-fe2020042219855

Esimerkkisyöttö OAMK:in Kirjasto- ja tietopalvelun koulutusohjelman opinnäytteellä: “Ammattikorkeakoulujen Julkaisuarkisto Theseus : käyttöönotto- ja perustamisvaiheet sekä nykyiset käytänteet ammattikorkeakouluissa” (2012)
https://www.theseus.fi/handle/10024/47634



Asiasanat: *

Asiasanaehdotukset – valitse sopivat

- [ ] kirjastot
- [ ] ammattikorkeakoulukirjastot
- [ ] Access
- [ ] sähköiset julkaisut
- [ ] kansalliskirjastot

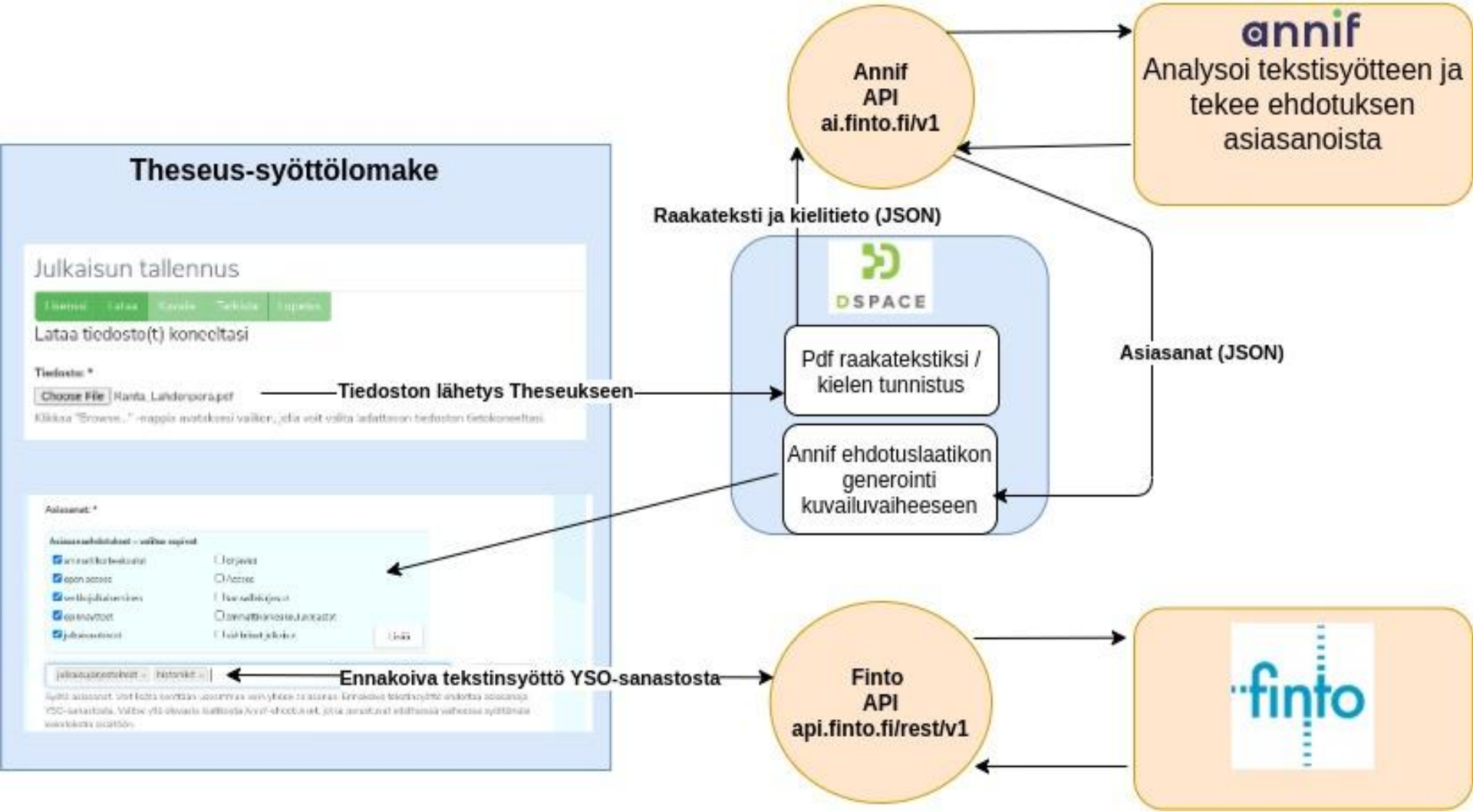
Lisää

Lisää

Syötä asiasanat. Voit lisätä kenttään useamman kuin yhden asiasanan. Ennakoiva tekstinsyöttö ehdottaa asiasanoja YSO-sanastosta. Valitse yllä olevasta laatikosta Annif-ehdotukset, jotka perustuvat edellisessä vaiheessa syöttämäsi kokotekstin sisältöön.

- [ ] ammattikorkeakoulut
- [ ] open access
- [ ] verkkojulkaiseminen
- [ ] opinnäytteet
- [ ] julkaisuarkistot
- [ ] julkaisujärjestelmät
- [ ] historiikit

Poista


Theseus-syöttölomake
Julkaisun tallennus
Lataa tiedosto(t) konseltasi
Tiedosto: *
Choose File
Kanta_Lahdenpera.pdf
Asiasanat: *
Tiedoston lähetys Theseukseen
Ennakoiva tekstinsyöttö YSO-sanastosta
Annif
API
ai.finto.fi/v1
Analysoi tekstisyötteen ja tekee ehdotuksen asiasanoista
Raakateksti ja kielitieto (JSON)
Asiasanat (JSON)
DSPACE
Pdf raakatekstiksi / kielen tunnistus
Annif ehdotuslaatikon generointi kuvailuvaiheeseen
Finto
API
api.finto.fi/rest/v1

Annifin ehdotusten dokumentointi

- Kaikki Annifin ehdottamat asiasanat ja yksilöivät URI.t tallennetaan piilossa oleviin kenttiin **dc.annif.suggestions** ja **dc.annif.suggestions.link**
    - Näkyvät vain admin-käyttäjille
- Mahdollistaa sen tutkimisen, miten Annifin valintoja on hyödynnetty
- Mahdollistaisi myös palauteväylän rakentamisen koneoppimisen tarpeisiin ehdotetuista vs. valituista sanoista
    - Palauteväylän rakentaminen kuitenkin vasta harkinnassa, ei välttämättä paranna tuloksia verrattuna jo käytössä oleviin ja tuleviin tekoälyalgoritmeihin

Puutteita ja paranneltavia asioita

- Tällä hetkellä Annif-syöttö Theseuksessa tukee vain yhden tiedoston syöttöä
  - Usean tiedoston syöttö ongelmallista, mm. kaikkien liitetiedostojen tekstit eivät välttämättä anna tukea asiasanoitukselle tai johtaa sitä harhaan
  - Ensimmäisenä syötetty teksti Annifiin yms. muita ratkaisuja harkitaan
- DC-metadata ei tue YSO-ontologian URI-tunnisteiden tallentamista
  - DC-muotoinen metadata ei mahdollista optimaalista tapaa yhdistää sanat ja URL:t toisiinsa.
  - Ei pelkästään syöttövaiheen ongelma:
    - Pitäisi myös ottaa huomioon URL:ien ja sanojen yhteys tietoja jäkeenpäin päivitettäessä / näytettäessä / tarjoiltaessa rajapintojen kautta

Kiitos!